Emotional content of an image attracts attention more than visually salient features in various signal-to-noise ratio conditions

Joanna Pilarczyk
Psychophysiology Laboratory, Institute of Psychology, Jagiellonian University, Krakow, Poland

Michał Kuniecki
Psychophysiology Laboratory, Institute of Psychology, Jagiellonian University, Krakow, Poland

Emotional images are processed in a prioritized manner, attracting attention almost immediately. In the present study we used eye tracking to reveal what type of features within neutral, positive, and negative images attract early visual attention: semantics, visual saliency, or their interaction. Semantic regions of interest were selected by observers, while visual saliency was determined using the Graph-Based Visual Saliency model. Images were transformed by adding pink noise in several proportions to be presented in a sequence of increasing and decreasing clarity. Locations of the first two fixations were analyzed. The results showed dominance of semantic features over visual saliency in attracting attention. This dominance was linearly related to the signal-to-noise ratio. Semantic regions were fixated more often in emotional images than in neutral ones, if signal-to-noise ratio was high enough to allow participants to comprehend the gist of a scene. Visual saliency on its own did not attract attention above chance, even in the case of pure noise images. Regions both visually salient and semantically relevant attracted a similar amount of fixation compared to semantic regions alone, or even more in the case of neutral pictures. Results provide evidence for fast and robust detection of semantically relevant features.

Introduction

Eye movements and attention

In humans, the area of acute color vision only spans around 2° in the center of the visual field, while the rest of the visual scene is blurry and hence lacking details. Therefore, in order to inspect the visual scene thoroughly, one has to execute a series of rapid and ballistic eye movements (saccades) intertwined with periods of gaze stability (fixations) during which the visual information is acquired. Under natural conditions, that is, when people spontaneously move their eyes, planning and executing saccades seem to closely reflect shifts of spatial visual attention (Rizzolatti, Riggio, Dascola, & Umiltá, 1987; Wright & Ward, 2008). This linkage allows one to study engagement of attention by using two alternative approaches in eye tracking: either examining properties of fixated regions or testing models predicting eye movements (for recent reviews see J. M. Henderson 2011, 2013; Schütz, Braun, & Gegenfurtner, 2011; B. W. Tatler, Hayhoe, Land, & Ballard, 2011). Particularly, the attentional mechanism involved in the detection of emotional stimuli may be examined using those methods. This is an interesting issue as several lines of experimental evidence converge with the conclusion that emotional visual stimuli are processed in a prioritized way. However, there is no agreement on the precise mechanism of this bias. Eye tracking may reveal how attention is deployed while viewing emotional scenes and thus provide insight into the process of discriminating between emotional and neutral objects.

Emotional images

Emotionally relevant scenes can be regarded as a special category of visual stimuli, particularly one that carries information critical for an organism's survival. Thus, tuning brain mechanisms to fast and prioritized processing of emotional stimuli seems to have an evolutionary adaptive value. Indeed, in a behavioral study, using the dot-probe paradigm, Mogg and

Citation: Pilarczyk, J., & Kuniecki, M. (2014). Emotional content of an image attracts attention more than visually salient features in various signal-to-noise ratio conditions. *Journal of Vision, 14*(12):4, 1–19, http://www.journalofvision.org/content/14/12/4, doi:10.1167/14.12.4.

doi: 10.1167/14.12.4 Received February 6, 2014; published October 7, 2014 ISSN 1534-7362

Bradley (1999) have shown that emotional images irrelevant to the experimental task are more likely to capture attention than neutral ones. Ohman, Flykt, and Esteves (2001) found an advantage of emotional over neutral stimuli in a task requiring subjects to find a designated target picture among a grid of distractors. Several studies employing electroencephalography found that the component of visual evoked potentials as early as P1 (100 ms after stimulus onset) or even C1 (80 and 100 ms after stimulus onset) differentiates between emotional and neutral stimuli (Carretié, Hinojosa, Martín-Loeches, Mercado, & Tapia, 2004; Pourtois, Grandjean, Sander, & Vuilleumier, 2004; Smith, Cacioppo, Larsen, & Chartrand, 2003; for a review see Olofsson, Nordin, Sequeira, & Polich, 2008), suggesting an automatic and ultrafast nature of this process. This conclusion is further strengthened by functional neuroimaging studies. Structures typically linked to the detection of emotional stimuli, such as the amygdala, are activated even when participants are not aware of an affective stimuli (Whalen et al., 1998) and regardless of whether participants' attention is directed to emotional stimuli or away from them (Vuilleumier, Armony, Driver, & Dolan, 2001). Also the eye-tracking data suggests preferential processing of emotional stimuli. Emotional images are more likely to first attract fixation when presented laterally together with the neutral image (M. G. Calvo & Lang, 2004; Nummenmaa, Hyönä, & Calvo, 2006), even if outside the focus of attention (M. Calvo, Nummenmaa, & Hyönä, 2007) or on the periphery of the visual field (M. Calvo, Nummenmaa, & Hyönä, 2008). However, it is worth noting that in those studies only a general bias towards the left or right visual field containing an emotional image was examined.

Features attracting fixations

The detection of emotionally relevant objects is an instance of the more general issue of how eye movements are guided towards any meaningful region of a visual scene. Fundamentally, fixation locations are not random; even the first fixation tends to fall in an informative region rather than in uniform areas of a scene (for a review see J. Henderson, 2003). However, it is not clear on what basis those regions are classified as informative by the nervous system. Two main groups of mechanisms have been proposed: analysis of low-level features or rapid extraction of semantic information.

Visual saliency

Low-level visual saliency of a scene is typically defined by local contrasts of luminance, opposing colors, or concentration of edges (Itti & Koch, 2000; Koch & Ullman, 1985). Visually salient regions carry more information than physically uniform ones, and are thus fixated more often (D. Parkhurst, Law, & Niebur, 2002). Salient physical features are supposed to influence attention in an automatic, bottom-up and stimulus-driven manner and are especially relevant at the early stages of image processing (J. M. Henderson, Weeks, & Hollingworth, 1999; Theeuwes, 2010; for a review see J. M. Henderson & Hollingworth, 1999). Theories advocating the role of those factors in guiding attention employ biologically plausible computational models to construct saliency maps of a visual scene, which is meant to predict consecutive fixation locations (Harel, Koch, & Perona, 2006; Itti & Koch, 2000; Koch & Ullman, 1985). Indeed, in the conditions of free exploration of the presented scenes, the first saccade usually falls on the areas with high contrast (D. J. Parkhurst & Niebur, 2003; Reinagel & Zador, 1999; B. W. Tatler, Baddeley, & Gilchrist, 2005), high concentration of edges (B. W. Tatler, Baddeley, & Vincent, 2006; B. W. Tatler et al., 2005), and, to a lesser degree, high chromaticity and luminance (B. W. Tatler et al., 2005).

Semantics

Attentional guidance based on semantics requires more elaborate information about a scene (e.g., identification of meaning of objects or emotional relevance of a scene). For the semantics to operate so early as to guide first fixations on a scene, at least a gist of the visual scene should be available to the nervous system almost immediately after the stimulus onset. Several studies provide support for the assumption that a gist is extracted before the first saccade is made. In a classic experiment, Potter (1975) found that participants were able to identify a target stimulus with nearly 80% accuracy in a constant stream of pictures each flashed for only 125 ms. Later it was shown that scene identification is possible even if the stimulus is flashed only for a split second not extending 30 ms and outside attentional focus (Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005; Kirchner & Thorpe, 2006; Li, VanRullen, Koch, & Perona, 2002; Peelen, Fei-Fei, & Kastner, 2009; Rousselet, Joubert, & Fabre-Thorpe, 2005; Thorpe, Fize, & Marlot, 1996).

Nevertheless, such rapid semantic analysis seems to be limited. Stimulus presentation lasting below 30 ms allows for only a crude gist determination without precise identification of the image (Greene & Oliva, 2009). To enable correct determination of image details, with 75% accuracy, the time of the presentation has to increase to about 50 ms. Also explicit categorization of a presented scene is possible no sooner than 150 ms after stimulus onset (Thorpe et al.,

1996; for a review see Hegdé, 2008). Rayner, Smith, Malcolm, and Henderson (2009) showed that if an experimental task requires finding a particular object in the presented scene, instead of just deciding upon the category to which it belongs, the minimum fixation time necessary to accomplish the task has to exceed 150 ms. Continuing this line of research, Võ and Henderson (2010) showed that preview of the scene lasting just 50 ms can already lead to a measurable temporal advantage in subsequent search task; however, to halve the search time, the preview time must be increased to 250 ms. Additional support for ultrafast categorization of semantic information influencing guidance of attention comes from the line of studies reporting that first fixations are more likely to be directed towards objects not matching the context (e.g., lawnmower in a kitchen) than ordinary ones (Brockmole & Henderson, 2008; Gordon 2004; Loftus & Mackworth, 1978). However, it has been questioned whether the effect of capturing first fixation by out-of-context objects really exists. Several reports show that odd objects are not fixated faster than regular ones, but once they are, the total fixation time tends to be longer (de Graef, Christiaens, & Ydewalle, 1990; Friedman & Liebelt, 1981; J. M. Henderson et al., 1999; Võ & Henderson, 2009, 2011; for review see Wu, Wick, & Pomplun, 2014).

Altogether, those findings suggest that eye movements can be influenced by meaning, since at least partial semantic analysis of the presented scene is conducted before the first saccade is executed. However, such analysis does not seem to be very thorough or complete at this early stage.

Interaction of semantics and visual saliency

Additionally, it is important to note that semantics are correlated with visual saliency, as meaningful objects tend to be salient (Einhäuser, Spain, & Perona, 2008; Elazary & Itti, 2008; J. M. Henderson, Brockmole, Castelhano, & Mack, 2007; J. M. Henderson, Malcolm, & Schandl, 2009; Schütz et al., 2011; B. W. Tatler et al., 2011). To determine which feature is dominant in guiding attention—saliency or meaning—Chen and Zelinsky (2006) investigated how semantic knowledge interacts with visual saliency in a search task. Participants had to find a simple character (+ or x) displayed over one of the of objects arranged on the screen in a circular shape. All of the objects on the display were gray except one, which was thereby made visually salient. On some trials the identity of the object containing the target character was revealed to the participant in a preview screen. They found that the visual salience played no role in attracting initial saccades, provided participants were primed to the possible target location in the preview condition. However, in the no-preview condition, visual saliency was clearly able to capture attention and attract initial saccades. Extending those findings to the perception of natural scenes, Foulsham and Underwood (2007) designed an experiment in which locations of visually salient regions and target objects were deliberately separated. When participants were allowed to freely explore the scene, visual saliency proved to influence both probability and speed of fixation. This relationship disappeared entirely once top-down bias was introduced by means of search task. Very similar design was used recently by J. M. Henderson et al. (2009), who showed that in a search task, nonsalient targets were fixated on average in over 90% of trials while salient regions only in around 10%. Einhäuser et al. (2008) used complex statistical modelling to prove that object maps predict fixation locations significantly better than saliency maps. Importantly, they established that variance accounted for by saliency is not unique; instead it can be largely explained away by semantic map. Very recently Onat, Açık, Schumann, and König (2014) contrasted user-defined interestingness maps for natural images against various visual saliency maps to determine the best predictor for fixations locations. Interestingness explained fixation locations much better than any of the saliency maps alone or any of their linear combination. Remarkably, the interestingness map was also effective in predicting fixations in case of fractal images, which bear no semantics but have object-like structures. Objects devoid of meaning do not belong either to a semantic or visually salient class. Therefore, it seems that the straightforward division between top-down semantic and bottom-up saliency factors in attracting fixations that has dominated the field may be oversimplified.

In summary, either object-based or semantic factors are dominant in guiding attention, especially in paradigms explicitly requiring participants to perform specific operation on visual stimulus (for review, see Wu et al., 2014). Nonetheless, in a free-viewing condition, saliency is capable of capturing attention; however, it might be due to its inherent entanglement with semantic or asemantic objects (Foulsham & Underwood, 2007; Koehler, Guo, Zhang, & Eckstein 2014; Onat et al., 2014; Schütz et al., 2011; B. W. Tatler et al., 2011; Underwood, Foulsham, van Loon, & Underwood, 2005).

Emotional meaning and visual saliency

Three recent studies addressed the question of which features attract visual attention while viewing emotional and neutral stimuli, yielding mixed results. In the first, Acunzo and Henderson (2011) transformed a set of innocuous images into negative, neutral, and positive

images by carefully planting an emotional object, be it a dangerous animal for negative, cute pet for positive, or a neutral object for neutral images, into the original scene. Looking at the average number of fixations preceding the one falling on an emotional object, they found no evidence that emotional content preferentially captures attention. However, once drawn to emotional content, attention was held longer than in the case of neutral objects. Humphrey, Underwood, and Lambert (2012) employed a very similar paradigm, but additionally they ensured that the visually salient region was located at a different spot than the emotional or neutral object planted in the scene. Their manipulation equalized the overall gist of the scenes and explicitly separated the salient from the affective regions. They established that the first fixation was more likely to fall within the affective region in the case of positive and negative images as compared to neutral ones. Similar results were reported by Niu, Todd, and Anderson (2012), who segmented pictures from neutral, negative, and positive categories into equally sized emotionally and physically salient regions. They found that the five initial fixations were more often directed towards affectively charged regions in the case of emotional pictures as compared to neutral ones. Additionally they established that the tendency to allocate first fixations in emotive rather than visually salient regions correlates linearly with arousal ratings of the picture.

Aims

The aim of the present study was to examine the dynamics of attention capture by semantics and visual saliency while viewing emotional images. Humphrey et al. (2012) and Niu et al. (2012) showed that semantically relevant locations almost completely monopolize first fixations, as compared to visually salient ones. We intended to examine how changes in clarity of a picture would influence this balance. This is an ecologically valid question since in natural conditions low visual clarity occurs rather often: Objects can be partly occluded or poorly lit. Detection of important objects, such as animals, is quite robust and remains unaffected even in the presence of large amounts of phase noise (Wichmann, Braun, & Gegenfurtner, 2006). It is not clear whether this property of the visual system includes detection of any meaningful objects, also neutral and uninteresting ones, or only those of emotional or behavioral significance. We hypothesized that low clarity should lead to an enhanced influence of visual saliency on eye movements, because in such an instance semantics is difficult to decipher while an image still possesses distinct salient regions. Our aim was to determine the threshold of the signal-to-noise ratio necessary for the semantic domination to appear and whether this threshold differs between emotional and neutral pictures.

Secondly, our goal was to explicitly investigate interactions between visual saliency and semantics in attention guidance. The design of previously described studies did not allow for examining this, as either an overlap of salient and emotional regions was deliberately prevented (Humphrey at al., 2012), or data from the overlapping regions were not analyzed (Acunzo & Henderson, 2011; Niu et al., 2012). In the present study, we did not prevent overlap of semantics and visual saliency, but treated it as a separate type of region of interest (ROI). Since both factors, semantics and visual saliency, are known to attract attention, we expected them to boost their ability to predict fixations. The optimal condition for studying interaction of these factors was free-viewing design, as it enables not only semantics but also saliency to influence fixations (Foulsham & Underwood, 2007; Koehler et al., 2014; Schütz et al., 2011; B. W. Tatler et al., 2011; Underwood et al., 2005).

An additional concern was the top-down influences on eye movements. Given our design involving multiple presentations of the image differing only in signal-to-noise ratio, the memory confound especially needed to be assessed. The influence of memory on eye movements is predicted by the scanpath theory (Noton & Stark, 1971), which assumes that a pictorial stimulus is memorized together with the sequence of fixations that accompanied its presentation. Once the picture is presented again, it invokes a pattern of memorized saccadic activity. It is currently doubted whether memory of visual stimulus indeed entails saccade reinstatement (Foulsham & Kingstone, 2013; J. Henderson, 2003). Instead it has been repeatedly shown that fixating the same locations in encoding and test phase improves recognition performance, while the specific order of fixations is not important. The same areas of the picture are revisited because people tend to fixate on informative or visually salient regions, which remain the same for every presentation of the stimulus (Foulsham et al., 2012; Foulsham & Kingstone, 2013; Foulsham & Underwood, 2008; Hollingworth & Henderson, 2002; Holm & Mäntylä, 2007; Tremblay, Saint-Aubin, & Jalbert, 2006). Controlling the top-down effects of memory was especially important due to the emotional content of displayed images, since extensive evidence has been accumulated for the differential influence of valence on memory formation (Cahill & McGaugh, 1998; S. Hamann, 2001; S. B. Hamann, Ely, Grafton, & Kilts, 1999; Humphrey et al., 2012; Humphreys, Underwood, & Chapman, 2010; LaBar & Cabeza, 2006; Packard & Cahill, 2001; Phelps, 2004). Additionally, in order to explicitly measure influence of valence on memory, we conducted a memory test shortly after the eye-tracking study.

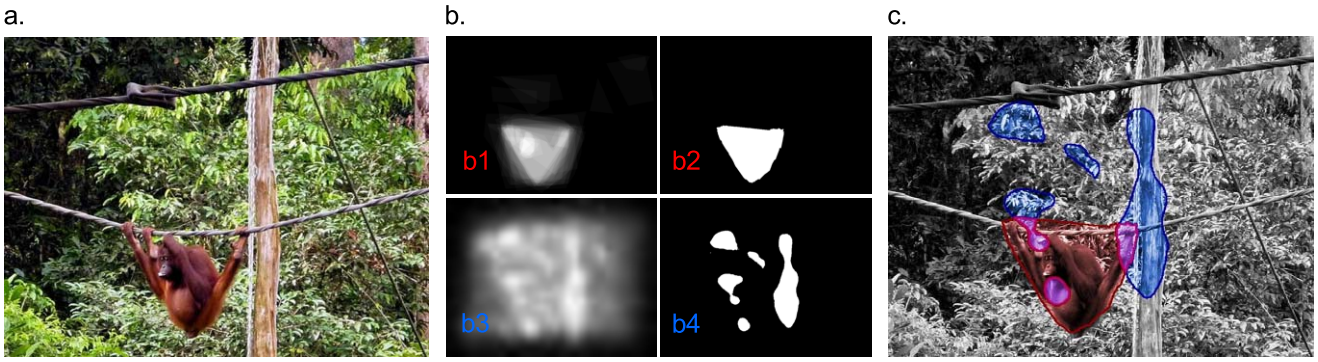

Figure 1. Examples of the original image (a), semantic map (b1), semantic ROIs (b2), saliency map (b3), and visually salient ROIs (b4). ROIs superimposed on original image (c) shows overlap (pink) of semantic ROIs (red) and visually salient ROIs (blue).

Materials and methods

Participants

Sixty-five participants (52 women), aged between 19 and 38 ($M = 21.6$) completed the semantic mapping task. Twenty students (14 women), aged between 19 and 25 ($M = 20.9$), with normal or corrected-to-normal vision and no history of neurological diseases participated in the eye-tracking experiment. Six of them were left-eye dominant. Twenty-seven participants (25 women), aged between 19 and 23 ($M = 20.6$), completed the classification task. All participants signed informed consent forms and received course credit.

Regions of interest

Semantic regions of interest

The goal of the construction of semantic maps was to determine regions of an image that are most relevant to the emotional (either positive, negative, or neutral) meaning of an image. A set of 120 pictures from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 2005) and Nencki Affective Picture System (NAPS; Marchewka, Żurawski, Jednoróg, & Grabowska, 2014) were used in the semantic mapping study (Figure 1a). Pictures were classified on the basis of the original IAPS and NAPS valence ratings as negative ($M = 2.47$, $SD = 0.73$), neutral ($M = 4.98$, $SD = 0.43$), or positive ($M = 6.99$, $SD = 0.51$). Pictures of sexual content were not included in the study as the valence and arousal ratings differ largely between men and women (Lang et al., 2005).

Participants were asked to circle key locations determining the valence of each picture using a simple computer tool. They had liberty to select as many regions on each picture as they chose, and of any size and shape, though they were also asked to keep it rather simple. Each participant marked ROIs on 60 images; hence, each image was rated by at least 30 participants. Every image was labeled with the valence category to avoid marking objects irrelevant to emotional content. Selections by all participants were averaged to construct a semantic saliency map (Figure 1 b1). Then, a threshold was applied to obtain regions marked by at least half of the participants (Figure 1 b2). The final set of 60 images used in the eye-tracking study was selected so that ROIs covered approximately 10% of an image (following Niu et al., 2012). ROI areas in the selected images did not differ between negative ($M = 9\%$), positive ($M = 9.1\%$), and neutral ($M = 9.1\%$) images; $F(2, 57) < 1$. However, the agreement among participants in labeling varied between valence categories, with the highest in case of negative (74%), lower in case of positive (65%), and the lowest in case of neutral images (56%); $F(2, 57) = 32.2$, $p < 0.001$. See Appendix A for estimation on how those differences impacted on the results.

Visually salient regions of interest

In order to localize visually salient regions, we generated saliency maps using the Graph-Based Visual Saliency (GBVS) model implemented by Harel et al. (2006; Figure 1 b3). The GBVS algorithm is a biologically plausible, bottom-up visual saliency model. It is based on similar assumptions as the classic Itti and Koch algorithm (2000), but proved to be more successful in predicting human fixations on natural images, both original and with modified contrast and luminance (Harel et al., 2006). The GBVS activation map is formed using graph computations, and the normalization step is designed to avoid uniform distribution of saliency and to highlight few informative locations. Therefore, GBVS maps structurally are rather similar to the semantic maps, with a few focused areas of interest.

The default GBVS saliency maps were processed in a similar manner to the semantic maps to obtain ROIs.

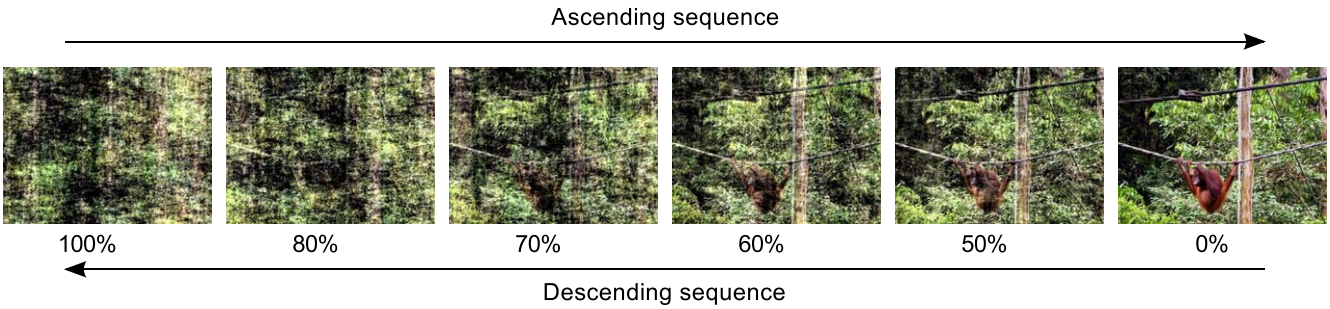

Figure 2. Experimental design. Five degraded versions and an original version of an image (percentage of pink noise content is indicated below each image) were presented in ascending and descending sequences, summing up to 11 images in each full sequence.

Namely, a threshold was applied to each GBVS map, so on each image the area of the visually salient region was equal to the area of the semantically salient region (Figure 1 b4).

Common ROIs

Some parts of the images ($M = 2.9\%$) were both semantically and visually salient, probably due to the fact that objects in general tend to be salient (Elazary & Itti, 2008). Those regions were analyzed separately from only semantically and only visually salient ROIs, comprising the third class of ROIs (common ROI; Figure 1c). This allowed us to examine the combined influence of semantic and salient features on attention allocation.

Pink noise

In order to manipulate signal-to-noise ratio, we mixed pink noise images with original images. Pink noise is obtained by replacing phase in the Fourier spectrum of an original image by random values between 0 and 2 pi while preserving the amplitudes (Kayser, Nielsen, & Logothetis, 2006). The inverse Fourier transformed pink noise image keeps some of the original image characteristics, like overall coloration, but all discriminable semantic information is lost.

For each original image, noise was added in the following proportions: 0, 50, 60, 70, 80, and 100%. During the eye-tracking study each image was displayed in ascending and descending sequence, i.e., from pure noise to complete lack of noise and back again to pure noise, forming a full sequence of 11 images (Figure 2). Importantly, the pink noise images were generated separately for each proportion, ascending and descending sequence, and each image. Therefore, no identical image was shown twice in the experiment. All images were equated (using Adobe Photoshop) in luminance and contrast, measured as mean and standard deviation of the L component in the L*a*b* color space. Then, a saliency map was computed for each image.

Eye-tracking study

Procedure

The eye-tracking study was divided into 10 blocks with self-paced breaks in between. In each block, six full sequences of images were presented. Each image was presented for 3 s, preceded by a central fixation cross displayed for 500 ms. The entire experimental procedure lasted up to 1 hr. Stimuli presentation was programmed using Eyelink Experiment Builder (SR Research, Ontario, Canada). Stimuli were presented on a 21-in. thin-film transistors (TFT) monitor. Each picture covered 23° of visual field in width and 17.5° in height. The monitor was calibrated using ColorMunki (X-Rite, Michigan, USA) to the white point CIE Illuminant D65 and luminance of 120 cd/m^2. Participants were seated 73 cm from the computer screen with their head position stabilized with a chinrest. Instructions given to the participants encouraged them to freely explore presented scenes with no task specified.

Eye movement recording and analysis

Eye position was recorded with an infrared remote Eyelink 1000 (SR Research) eye tracker, sampling pupil position at 500 Hz. Position of both eyes was recorded, but only the data from the better-calibrated eye was analyzed. A 9-point calibration and validation procedure was repeated at the beginning of each experimental block and whenever necessary. Average calibration error of the analyzed eye was 0.39° ($SD = 0.12$). Fixations and saccades were detected using default Eyelink 1000 algorithm. Saccades were defined as deflections greater than 0.15°, of velocity exceeding 30°/s and of acceleration over 8000°/s^2. Fixations were defined as periods between saccadic eye movements. Although we did not set duration threshold to

eliminate microfixations, fixations longer than 100 ms comprised 96% of the analyzed data.

The locations of the first two fixations on each image were analyzed by whether they fell into one of the three kinds of ROIs: semantic, visually salient, or common ones. The first fixation was defined as starting after the onset of an image. M. Calvo et al. (2008) showed that eye movements up to 500 ms after the stimulus onset are out of voluntary control. Thus, narrowing our analysis to only the first two fixations allowed us to study eye movements guided by bottom-up processes rather than volitional ones.

Number of fixations falling within a ROI depends not only on the properties of this particular area but also on size and position of a ROI; for example, due to the tendency to look at the center of the screen (D. Parkhurst et al., 2002; B. Tatler, 2007; B. W. Tatler, Baddeley, & Gilchrist, 2005). To exclude such confounds, we computed the fraction of fixations that fall within a ROI while looking at the analyzed image (positive sample), and the fraction of fixations made in this location in response to different images (negative sample). Size of the negative sample indicates a tendency to look at this particular location of a scene regardless of the location's features. Percent values of the positive sample were divided by the percent value of the negative sample, creating a normalized index of fixation proportion, similar in principles to other tests of classificator's strength like commonly used receiver operating curve (ROC; see Appendix B for direct comparison of normalized fixation proportion [*NFP*] measure with ROC). If *NFP* equals one, number of fixations in a ROI can be explained by the ROI's location and size. If it is larger, more fixations fall in a ROI than predicted by chance. If the *NFP* index is smaller than one, a ROI receives a smaller proportion of fixations than predicted by its size and position. It is worth noting that this approach compensates for the size of a ROI and participants' tendencies to scan central areas of an image more often than peripheries; however, it does not take into account photographers' tendency to place objects in the center of a photo. Thus, it may underestimate attraction of attention by centrally located objects. For a comprehensive review of this issue, see Tatler et al. (2005).

Memory

Descending sequence

The descending sequence served as a test for memory influence on attention deployment. The memory effects could have occurred only in the case of semantic ROIs, since the semantic map remained constant for all images across the full sequence, while the visual saliency map varied for every image. If the descending sequence for semantic ROIs produced a curve symmetric to the ascending one, we would assume that top-down influences were weak and that the eye movements were guided by stimuli presented on the screen at the very moment. However, if fixations were linked to memory of the semantic ROIs' locations, participants should still examine those regions in the descending sequence. This should lead to relative immunity from noise interference and hence cause a rise of values and flattening of the curve in descending sequence. Controlling for memory effects was the main reason for sequential design rather than more common random presentation, as the latter would not allow for comparing ascending with descending sequences.

Memory task

Additionally, a recognition test was conducted 10 min after the eye-tracking study. A set of 120 images was presented with the instructions to decide whether each image had been displayed earlier in the eye-tracking study (a yes-no decision). Half of the images in the set were new, and the other half comprised all of the original images from the eye-tracking study. New images were matched in valence and arousal ratings with the old set.

Classification task

In a complementary study, we investigated how noise impacted subjects' ability to identify the gist of the image by asking them to judge whether the presented image depicts an outside or inside scene. In this computer task, the same pictorial material was used as in the eye-tracking study, with the same timing parameters, except that the descending sequence was omitted. Images spanning 23.1° in horizontal and 18.3° in vertical planes were presented on a 20-in. TFT monitor located 60 cm from the participant. The monitor was calibrated to D65 white point and luminance equal to 120 cd/m^2. Ascending sequences (Figure 2) were presented in a random order. After the presentation of each image, a response screen appeared prompting the participant to make a judgment of whether the presented scene was taken outside or in the interior.

Results

Regions of interest

To investigate the relationship between attention and emotional content of the picture across changing

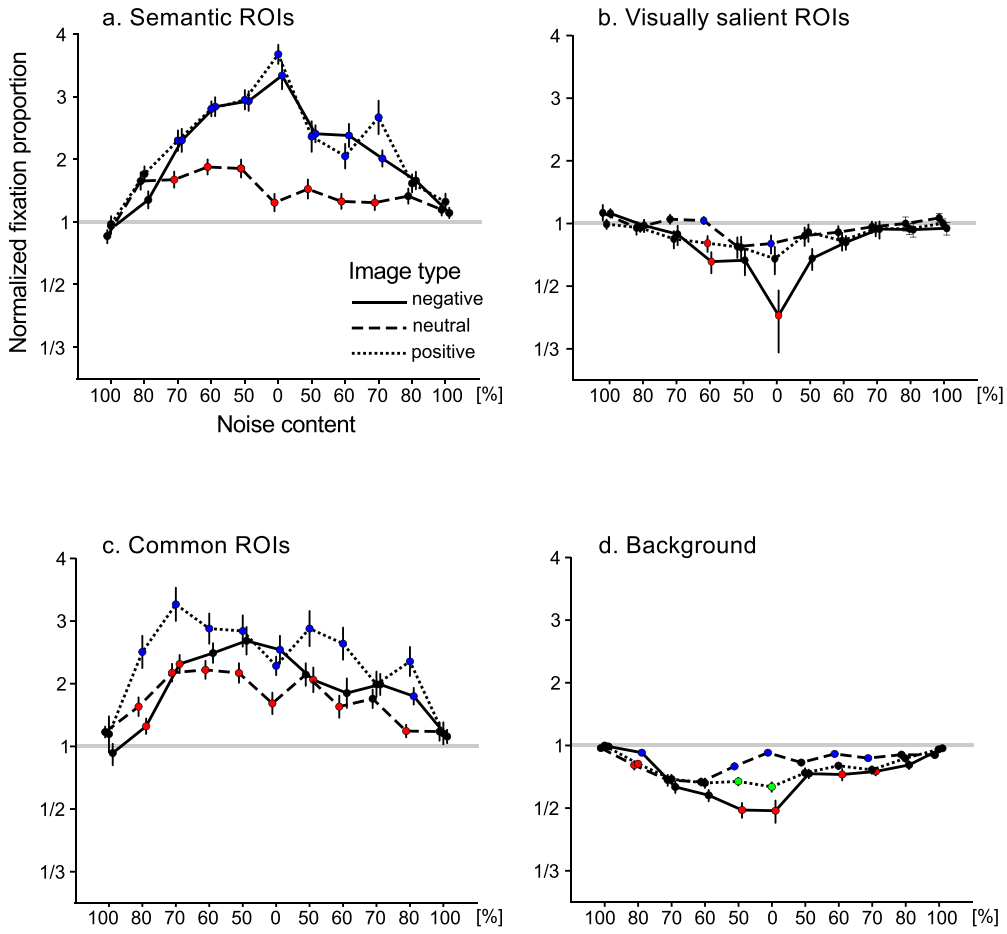

Figure 3. *NFP* for each ROI: semantically rich but not visually salient (a), visually salient but not semantically rich (b), both visually salient and semantically rich, dubbed "common ROIs" (c), and neither visually salient, nor semantically rich, dubbed "background" (d). The gray line represents the level on which observed fixation proportion is equal to proportion explained by region size and location. On the *x*-axis, percentage of noise content in an image is indicated, arranged in the same order as presented in the experiment. Statistically significant differences between emotional conditions are encoded by color of dots (blue, red, and green) representing data points. If a pair differed significantly, corresponding dots have different colors. If a condition did not differ from any other, the dot is black. Error bars represent standard deviations.

levels of visibility, we used a repeated measures ANOVA design with factors of valence (negative, neutral, positive) and noise (11 levels). In order to explore not only overall differences between the group means but also differences in the shapes of curves describing the relationship between noise levels and chances of attracting fixations, omnibus ANOVAs were complemented by analysis of linear, quadratic, and cubic trends. If several trends were simultaneously significant, we report the one explaining the largest proportion of variance. In all cases in which the sphericity assumption has been violated, the results are reported with H-F correction. Simple effects were investigated using Bonferroni correction. One participant was excluded from the analysis due to large data loss (30.2 %) as compared to average (3.9 %, $SD = 4.4$).

Semantic ROIs

Fixations fell more often in semantically rich areas (Figure 3a) than in visually salient ones (Figure 3b). Specifically, semantic ROIs attracted almost two times more fixations than expected by chance ($NFP = 1.9$), while visually salient ROIs less than expected ($NFP = 0.86$).

The chance to fixate in semantic ROIs was modulated by the level of pink noise, $F(10, 180) = 46.2$, $p < 0.001$. This effect was stronger in the case of emotional than neutral pictures, $F(2, 36) = 87.4$, $p < 0.001$. Also an interactive effect of valence and signal-to-noise ratio emerged, $F(20, 360) = 7.8$, $p < 0.001$. Analysis with orthogonal polynomials showed that the relationship between the noise level and the chance of attracting fixation by semantically rich regions can be explained best by the quadratic trend, $F(1, 18) = 291.9$, $p < 0.001$.

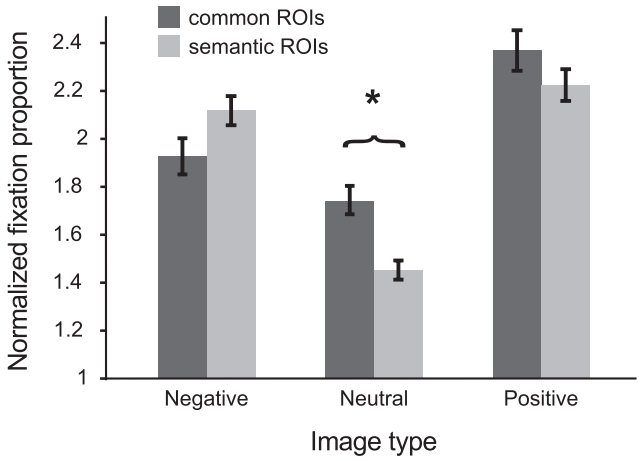

Figure 4. Normalized fixation proportions for semantic ROIs and common ROIs, averaged across all noise conditions. Statistically significant difference is marked with an asterisk. Error bars represent standard deviations.

Further investigation of this effect with planned contrasts revealed that parabolic relationship between the noise level and likelihood of attracting initial saccades differs between emotional and neutral conditions, $F(1, 18) = 99.1$, $p < 0.001$. Fixation chance starts differentiating between neutral and emotional conditions at 70% of noise and continues to be higher, until passing the level of 70% of noise on the descending sequence.

Visually salient ROIs

Also in the case of visually salient regions, signal-to-noise ratio influenced attraction of attention, $F(10, 180) = 10.6$, $p < 0.001$, as did emotional category of the stimulus, $F(2, 36) = 4.9$, $p = 0.013$. The relationship between level of noise and the chance of attracting initial saccade proved to be quadratic, $F(1, 18) = 62.7$, $p < 0.001$. The chance of attracting attention by visually salient ROIs starts at the chance level and decreases as a function of picture visibility. Planned orthogonal contrasts showed that the more concave curve in the negative condition is indeed significantly different than flatter ones in neutral and positive conditions, $F(1, 18) = 5.7$, $p = 0.028$.

Common ROIs

ROIs both semantically rich and visually salient (or common ROIs for short) were almost as good at attracting initial fixations as purely semantic ones (Figure 3c). On average they attracted two times more fixations than expected by chance. Also the pattern of results was very similar to the one observed with semantic ROIs. The noise level as well as the valence of the picture were both highly significant in modulating attraction of attention by common ROIs, $F(10, 180) = 26.7$, $p < 0.001$ and $F(2, 36) = 24.4$, $p < 0.001$, respectively. The interaction of valence and noise level was significant as well, $F(20, 360) = 2.7$, $p < 0.001$. The shape of the curve describing the relationship between the noise level and the chance of the common ROIs to attract fixation proved to be best fit by a parabolic curve, $F(1, 18) = 122.8$, $p < 0.001$. Interaction of orthogonal contrasts with this quadratic trend, $F(1, 18) = 14.6$, $p = 0.001$, proved that the curve in the neutral condition was different as compared to emotional conditions.

In order to directly compare the potential for attracting attention between semantic ROIs and common ROIs, an additional ANOVA was conducted with factors of ROI type (semantic vs. common), valence (negative, neutral, positive), and noise (eleven levels). The interaction of valence and ROI type was highly significant, $F(2, 36) = 9.3$, $p = 0.001$. Follow up investigation of this interaction with simple effects revealed that in the neutral condition common ROIs attracted more fixations than semantic ROIs ($p = 0.002$), while there was no difference for negative and positive conditions (Figure 4).

Background

Background was defined as the remaining image area, which was neither semantically rich nor visually salient. The chance of an early fixation falling in the background was lower than for any type of analyzed ROIs ($NFP = 0.75$; Figure 3d). The chance of fixating in the background depended on both noise level, $F(10, 180) = 47.2$, $p < 0.001$, and valence of the picture, $F(2, 36) = 18.3$, $p < 0.001$. Also the interaction of those two factors was significant, $F(20, 360) = 7.6$, $p < 0.001$. Examination with orthogonal polynomials showed that the curve representing the relationship between the chance of fixation and noise level follows a quadratic trend, $F(1, 18) = 173$, $p < 0.001$. Furthermore, this parabolic trend differed between the valence conditions, as revealed by planned linear, $F(1, 18) = 8.3$, $p = 0.010$, and quadratic, $F(1, 18) = 42.2$, $p < 0.001$, contrasts.

Memory

Two complementary methods were used to assess possible memory influence on fixation location: comparing ascending to descending sequence and memory task. To compare sequences, we performed three separate repeated measures ANOVAs, one for each valence category, with factors of sequence (ascending, descending) and noise intensity (five levels). The descending sequence differed from ascending only in

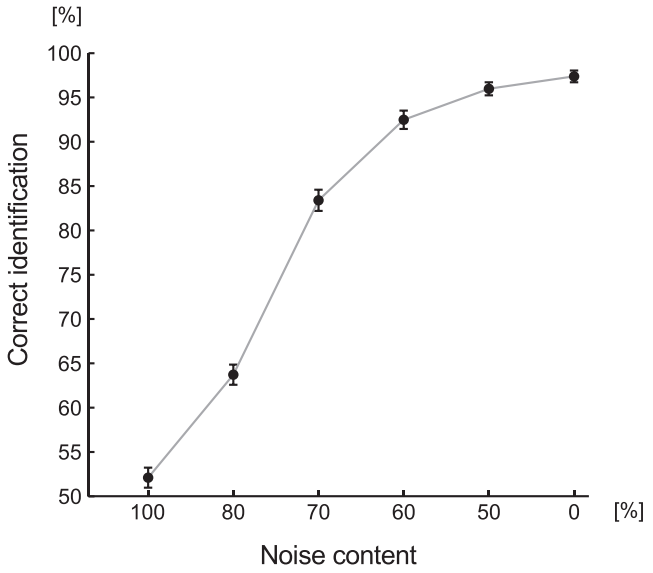

Figure 5. Percent of correct identifications of scene type (open space or interior) for images with different percent of noise. Random classification resulted in 50% of correct responses. Error bars represent standard deviations.

the case of neutral pictures with values in descending sequence markedly lower than in the ascending one. This effect was validated by the interaction of a tendency to fixate within semantically rich region and level of noise, $F(4, 72) = 2.7$, $p = 0.036$. Differences for negative and positive pictures did not reach significance. In the memory task, the probability of correctly discerning old and new pictures was highest in the negative condition (97%) and lower in neutral (94%) and positive (92%) conditions. This might point to the potential memory advantage of negative scenes. However, the level of false alarms (that is percentage of new pictures incorrectly recognized as old) was also largest for negative as compared to neutral and positive pictures (4%, 2%, and 2%, respectively). Consequently, d' measure, which takes into account both correct answers and false alarms, did not differentiate significantly between the valences, $F(2, 38) = 0.48$.

Classification task

The relationship between the accuracy of scene identification and the noise level is depicted in Figure 5. As expected, in the pure noise condition participants were not able to identify scene setting better than chance (50%). Decreasing noise level to 80% allowed participants to achieve 63.7% accuracy; decreasing it further to 70% resulted in the largest gain in accuracy to the level of 83% of correct answers. Subsequently, the accuracy of responses increased linearly with the decrease of noise, reaching 97% when the picture was fully visible.

Discussion

The aim of this experiment was to investigate early stages of visual attention deployment to emotional and neutral content obscured by asemantic noise, by analyzing locations of the first two fixations. Two types of features were taken into account as potential attractors of attention: low-level visual saliency and semantic relevance, as well as their interaction. Secondly, the balance of the number of fixations guided by saliency versus guided by semantics was examined under fluctuating conditions of low and high clarity of stimuli. In addition, the influence of memory on fixations was assessed to control for top-down confounds caused by sequential presentations of images differing only in the noise ratio.

Semantics

Results regarding the relative influence of low-level features and emotional relevance on fixations are in concordance with two earlier studies (Humphrey et al., 2012; Niu et al., 2012). Specifically, while viewing clearly visible images (0% noise condition), subjects were more likely to fixate on semantic ROIs than on visually salient ones. In fact, semantic superiority effect was so robust that it was visible across all but pure noise conditions. This domination of semantic factors in attracting initial fixations was more pronounced in the case of emotional than neutral images, effect reported also by M. G. Calvo and Lang (2004), M. Calvo et al. (2007, 2008), Humphrey et al. (2012), and Nummenmaa et al. (2006), but see Acunzo and Henderson (2011) for contradicting results. Since the tendency to fixate in semantic ROIs already emerged when semantic information was available for the first time in the sequence, we can assume that at least crude semantic analysis is indeed executed swiftly enough to alter initial fixations, which corroborates previous findings (Chen & Zelinsky, 2006; Humphrey et al., 2012; Kayser et al., 2006; Kirchner & Thorpe, 2006; Võ & Henderson, 2010). Our data show that this effect persists even under conditions of low visibility such as 80% of noise. However, at this level of noise no discrimination between emotional and neutral images appeared, suggesting general process of object detection rather than analysis of emotional meaning. This observation is further supported by examination of identification curve (Figure 5) which shows that at 80% of noise participants were able to discern scene content

with 63% of accuracy (i.e., only 13% above chance level). At 70% of noise condition the accuracy ratio rose sharply to 83%. Interestingly, at the same level of noise, the tendency to fixate in semantically rich regions started to differentiate between neutral and emotional conditions. This suggests that also attentional dominance of emotional stimuli could be explained in terms of fast and robust semantic analysis. This conclusion is further supported by Schupp et al. (2008), who conducted an EEG experiment exploring the influence of random noise on processing of emotional stimuli with Early Posterior Negativity (EPN), known to vary with level of arousal induced by emotional stimuli (Junghöfer, Bradley, Elbert, & Lang, 2001). They found, just like in the present study, that at 70% of noise the accuracy of stimuli identification markedly increased above a random level. Importantly, the difference in EPN between high and low arousing stimuli started to emerge at exactly the same noise condition. Schupp and colleagues (2008) concluded that by the time EPN peaks (200–300 ms), the stimulus meaning is already identified, and therefore the component occurs as long as low-level features are not obscured by noise beyond recognition.

Visual saliency

Low-level visual saliency free from any semantics does not predict fixations above chance. Lack of predictive power of saliency was observed even at the beginning of the sequence, when the presented stimulus comprised pure noise. Poor performance of the GBVS model may be somewhat surprising, as it proved to be effective at predicting fixation locations (Emami & Hoberock, 2013; Harel et al., 2006). However, the model performed considerably better if predicted locations were also semantically relevant, as common ROIs were attracting a similar share of fixations as purely semantic ones. This observation leads to the conclusion that meaning is the driving force behind the tendency to fixate in common ROIs, which was previously postulated by Einhäuser et al. (2008) and J. M. Henderson et al. (2009). Furthermore, this pattern of results indicates that the previously reported predictive power of the GBVS model (and probably also similar ones) is based mainly on the correlation of low-level salient features with semantics. Overestimation of visual saliency influence on eye movements caused by correlation with semantics has been also suggested by J. M. Henderson et al. (2007). They showed that the fixated regions differed from non-fixated both in terms of contrast, intensity, and edge density as well as semantic content.

The strength of the correlation between salience and meaning depends on the image content: social scenes, particularly faces and eye regions, do not tend to be salient, while everyday objects do (Elazary & Itti, 2008). Accordingly, visual saliency models' predictive power depends on the image category, with particularly poor performance in case of social scenes (Birmingham, Bischof, & Kingstone, 2009) and good performance in case of inanimate scenes (D. Parkhurst et al., 2002). Therefore, it seems that if there are no semantic features with which saliency might correlate, the saliency model fails at predicting fixations. Our results suggest that under specific conditions visual saliency can increase predictive power of semantics. In the neutral category more fixations were drawn to overlap of semantic and salient regions than to purely semantic ones (Figure 4). Presumably because objects in semantic neutral ROIs do not convey important message, it is not vital to analyze them in detail. Thus the spectator of a neutral scene is more prone to dwell over brighter or higher contrast areas of the objects.

The relationship between clarity of the picture and the magnitude of semantic dominance in attracting fixations points to sharp differences between neutral and emotional pictures. In the case of negative and positive stimuli, the tendency to fixate in semantic regions was linearly related to the legibility of the picture. This linear relationship was stable across all noise levels up to the no-noise condition, while in the case of neutral stimuli it broke down after reaching a threshold of 80% of noise (see Figure 3a). Interestingly, the chance of attracting fixations by neutral semantic ROIs started to diminish already after passing 60% of noise. Similar differences between emotional and neutral pictures were visible in the case of common ROIs. While emotional common ROIs attracted more fixations as their clarity grew, the tendency to fixate in neutral common ROIs rose up to reach its maximum in 60% of noise and declined afterwards (Figure 3c). This decrease was not accompanied by an elevated tendency to fixate in visually salient ROIs; hence, the loss in tendency to fixate on semantic regions cannot be explained by competition from visually salient ROIs. Instead, background—that is, areas neither semantically rich nor visually salient—was fixated on more often. Conversely, in the case of positive and negative images, the probability of fixating on background continued to decline as the noise level decreased down to the point of the noise free condition. Taking into account that at 60% of noise the identity of the presented image was clear to participants, it seems that after recognizing visual stimuli as innocuous, they lost interest in exploring them in more detail. Objects in neutral pictures presumably do not differ in emotional meaning from the background as much as emotional objects do. As argued by Acunzo and Henderson (2011) emotional objects can be compared to gist-inconsistent while neutral to gist-consistent ones. Thus,

in case of neutral images, semantic meaning of the background was of similar value to meaning of the object, and fixations were distributed more evenly. In contrast, in emotional images the difference between object and background was of qualitative nature, causing more narrow focusing of attention.

Sustained tendency to attend emotional objects might reflect the way the brain tags possibly important stimuli. Several researchers (Davis & Whalen, 2001; Pessoa, 2011; Sander, Grafman, & Zalla, 2003; Whalen, 1998) postulated that the amygdala, a small almond-shaped subcortical structure, is specially poised to quickly discern mundane from potentially significant stimuli. The amygdala receives connections from higher-level temporal visual cortex, enabling transfer of object information. It also sends projections back to the visual areas, including V1, enabling sensitization of visual cortex in case an important object has been detected (Freese & Amaral, 2005; Sabatinelli, Lang, Bradley, Costa, & Keil, 2009). Our results allow us to speculate that the amygdala was active throughout the whole ascending sequence of emotional stimulus emergence, boosting the visual system to extract more details. In contrast, neutral content could have silenced the amygdala at the very early phases of emergence, as even the rudimentary information available at this stage was enough to determine that the perceived stimulus was not worthy of further investigation.

Memory

The hypothesized memory effect was expected to manifest as deviation from symmetry between two parts of the curve related to the ascending and descending sequence for semantic ROIs. Both negative and positive images produced symmetric curves, suggesting that the memory of fixated regions in ascending sequence did not influence the pattern of fixations in the descending sequence. Therefore, in case of emotional images, saccades were elicited by the bottom-up processes related to the presentation of the stimulus, rather than by the top-down effects linked to memory of semantically relevant regions. Similar conclusions were reached by Foulsham and Kingstone (2013) who studied relationship between scanpath and memory. They found that accuracy in the recognition phase did not differ regardless of whether participants were shown regions fixated in encoding phase by themselves or by other participants. Thus, it is rather the informative content of the image that determines fixation locations than top-down effect such as oculomotor memory trace.

The asymmetry was observed only for neutral images. As mentioned earlier, in the case of neutral images, the ascending sequence produced a parabolic curve reaching its apex at 60% noise and dropping afterwards to reach almost random level in the no noise condition. The entire descending sequence did not deviate from this level, producing almost a flat line. The direction of the asymmetry was opposite from what was expected, as the descending sequence was flattened but with lowered instead of elevated overall values. Thus, the memory effect for the presented neutral scenes, although present, did not result from revisiting previously fixated locations. Instead, the memory effect we detected was more global and possibly linked to the classification of the stimulus as unworthy of closer examination, as lowering of *NFP* to semantic ROIs was mirrored by rise in *NFP* to background ROIs.

Lack of top-down memory effects in fixation tendency corresponds with the results of control task designed to check for the memory advantage of emotionally charged pictures. In this task participants were equally good at discerning old from new pictures regardless of their valence. This is somewhat surprising, taking into account a large body of research pointing to the memory advantage of the emotional over neutral stimuli. In our case, the memory superiority effect for emotional stimuli might have been prevented by overlearning. Every picture in our study was shown in sequence of 11 expositions, leading to a ceiling effect in the memory task. The ceiling effect might have been alleviated by presenting images in random order. However, as mentioned before, random design would not allow for comparing descending and ascending sequences, thus preventing more precise measure of memory effect. Moreover, our results suggest that that attention is guided chiefly by features of currently presented stimulus and not by its memory, and thus in case of random presentation, the observed effects of visual saliency and semantics on fixation distribution would be similar.

In conclusion, it seems that semantic domination over visual saliency cannot be attributed to memory effects. Thus, the obtained pattern of results points to analysis of semantic content of a scene as the main factor guiding first saccades. Such analysis needs to be quite complex, robust to noise interference and, above all, ultrafast. However, there might be an alternative explanation of the phenomenon of ultrafast differentiation of semantic scenes. It would involve analysis of higher-order statistical parameters of the presented picture, which could serve as a heuristic for image classification. Indeed, in some cases higher-order statistics satisfactorily explain differences between images. For example, fast recognition of face images can result from analysis of differences in spectrum of amplitudes of Fourier components (Honey, Kirchner, & VanRullen, 2008), whereas differentiating animals from other objects can rely on interaction between the amplitude and the phase of their Fourier components

(Gaspar & Rousselet, 2009). Moreover, Krieger, Rentschler, Hauske, Schill, and Zetzsche (2000) claim that attracting attentional focus can be explained by differences in spatial frequency bispectrum, which synthetically reflects such image features such as the presence of curved lines, edges, T-junctions, isolated objects, and occlusions. Physical features may thus form the basis even for complex image classification, but analysis of higher-order physical features seem to be considerably different from simple visual saliency detection, bearing more resemblance to semantic object recognition.

Conclusions

In summary, the present experiment provides strong support for the notion that saccades are guided by semantics and not by visual saliency. Visual saliency alone might play some minor role in attracting fixations in the case of "weak semantics," i.e., neutral stimuli. Correlation of semantics with visual saliency may account for effective fixation predictions made by visual saliency models. Semantic differentiation of stimuli between emotional and neutral is achieved fast and is robust against distortion. Stimuli identified as emotional attract and sustain attention, while those identified as neutral are ignored. Top-down factors such as spatial memory have negligible influence on eye movements in comparison to bottom-up factors related to the ongoing stimulus presentation.

Keywords: eye tracking, emotion, visual saliency, attention, signal-to-noise ratio

Acknowledgments

This work was supported by the Polish National Science Centre [grant number N N106 288939 and 2012/07/E/HS6/01046].

Commercial relationships: none.
Corresponding author: Michał J. Kuniecki.
Email: michal.kuniecki@uj.edu.pl.
Address: Psychophysiology Laboratory, Institute of Psychology, Jagiellonian University, Krakow, Poland.

References

Acunzo, D. J., & Henderson, J. M. (2011). No emotional "pop-out" effect in natural scene viewing. *Emotion*, *11*(5), 1134–1143.

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research*, *45*(11), 1459–1469.

Birmingham, E., Bischof, W., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, *49*(24), 2992–3000.

Brockmole, J. R., & Henderson, J. M. (2008). Prioritizing new objects for eye fixation in real-world scenes: Effects of object–scene consistency. *Visual Cognition*, *16*(2–3), 375–390.

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences*, *21*(7), 294–299.

Calvo, M., Nummenmaa, L., & Hyönä, J. (2007). Emotional and neutral scenes in competition: Orienting, efficiency, and identification. *Quarterly Journal of Experimental Psychology*, *60*(12), 1585–1593.

Calvo, M., Nummenmaa, L., & Hyönä, J. (2008). Emotional scenes in peripheral vision: Selective orienting and gist processing, but not content identification. *Emotion*, *8*(1), 68–80.

Calvo, M. G., & Lang, P. J. (2004). Gaze patterns when looking at emotional pictures: Motivationally biased attention. *Motivation & Emotion*, *28*(3), 221–243.

Carretié, L., Hinojosa, J. A, Martín-Loeches, M., Mercado, F., & Tapia, M. (2004). Automatic attention to emotional stimuli: Neural correlates. *Human Brain Mapping*, *22*(4), 290–299.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, *46*(24), 4118–4133.

Davis, M., & Whalen, P. J. (2001). The amygdala: Vigilance and emotion. *Molecular Psychiatry*, *6*(1), 13–34.

DeGraef, P., Christiaens, D., & Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*, 317–329.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14):18, 1–26, http://www.journalofvision.org/content/8/14/18, doi:10.1167/8.14.18. [PubMed] [Article]

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3):3, 1–15, http://www.journalofvision.org/content/8/3/3/, doi:10.1167/8.3.3. [PubMed] [Article]

Emami, M., & Hoberock, L. L. (2013). Selection of a best metric and evaluation of bottom-up visual

saliency models. *Image & Vision Computing*, *31*(10), 796–808.

Foulsham, T., Dewhurst, R., Nyström, M., Jarodzka, H., Johansson, R., Underwood, G., & Holmqvist, K. (2012). Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach. *Journal of Eye Movement Research*, *5*(4), 1–14.

Foulsham, T., & Kingstone, A. (2013). Fixation-dependent memory for natural scenes: An experimental test of scanpath theory. *Journal of Experimental Psychology. General*, *142*(1), 41–56.

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception*, *36*(8), 1123–1138.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2):6, 1–17, http://www.journalofvision.org/content/8/2/6/, doi:10.1167/8.2.6. [PubMed] [Article]

Freese, J. L., & Amaral, D. G. (2005). The organization of projections from the amygdala to visual cortical areas TE and V1 in the macaque monkey. *Journal of Comparative Neurology*, *486*(4), 295–317.

Fredman, A., & Liebelt, L. S. (1981). On the time course of viewing pictures with a view towards remembering. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception* (pp. 137–155). Hillsdale, NJ: Erlbaum.

Gaspar, C., & Rousselet, G. (2009). How do amplitude spectra influence rapid animal detection? *Vision Research*, *49*, 3001–3012.

Gordon, R. D. (2004). Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception & Performance*, *30*(4), 760–777.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176.

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences*, *5*(9), 394–400.

Hamann, S. B., Ely, T. D., Grafton, S. T., & Kilts, C. D. (1999). Amygdala activity related to enhanced memory for pleasant and aversive stimuli. *Nature Neuroscience*, *2*(3), 289–293.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Proceedings of Neural Information Processing Systems (NIPS)*.

Hegdé, J. (2008). Time course of visual perception: coarse-to-fine processing and beyond. *Progress in Neurobiology*, *84*(4), 405–439.

Henderson, J. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

Henderson, J. M. (2011). Eye movements and scene perception. In I. D. Gilchrist, S. P. Liversedge, & S. Everling (Eds.), *Oxford handbook of eye movements* (pp. 593–606). Oxford: Oxford University Press.

Henderson, J. M. (2013). Eye movements. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 69–82) New York: Oxford University Press.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. L. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. L. van Gompel, R. P. G. Fischer, M. H., Murray, & W. S., Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Oxford, UK: Elsevier Ltd.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–71.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856.

Henderson, J. M., Weeks, P. A., Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception & Performance*, *25*(1), 210–228.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception & Performance*, *28*(1), 113–136.

Holm, L., & Mäntylä, T. (2007). Memory for scenes: Refixations reflect retrieval. *Memory & Cognition*, *35*(7), 1664–1674.

Honey, C., Kirchner, H., & VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *Journal of Vision*, *8*(12):9, 1–13, http://www.journalofvision.org/content/8/12/9/, doi:10.1167/8.12.9. [PubMed] [Article]

Humphrey, K., Underwood, G., & Lambert, T. (2012). Salience of the lambs: A test of the saliency map hypothesis with pictures of emotive objects. *Journal*

of Vision, *12*(1):22, 1–15, http://www.journalofvision.org/content/12/1/22, doi:10.1167/12.1.22. [PubMed] [Article]

Humphreys, L., Underwood, G., & Chapman, P. (2010). Enhanced memory for emotional pictures: A product of increased attention to affective stimuli? *European Journal of Cognitive Psychology*, *22*(8), 1235–1247.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506.

Junghöfer, M., Bradley, M. M., Elbert, T. R., & Lang, P. J. (2001). Fleeting images: A new look at early emotion discrimination. *Psychophysiology*, *38*(2), 175–178.

Kayser, C., Nielsen, K. J., & Logothetis, N. K. (2006). Fixations in natural scenes: Interaction of image structure and image content. *Vision Research*, *46*(16), 2535–2545.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*, 1762–1776.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*(4), 219–227.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision*, *14*(3):14, 1–27, http://www.journalofvision.org/content/14/3/14, doi:10.1167/14.3.14. [PubMed] [Article]

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision*, *13*(2–3), 201–214.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7*(1), 54–64.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2005). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual* (Technical Report A-8). Gainesville, FL: University of Florida.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, USA*, *99*(14), 9596–9601.

Loftus, G., & Mackworth, N. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception & Performance*, *4*(4), 565–572.

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic pictures database. *Behavior Research Methods*, *46*(2), 596–610.

Mogg, K., & Bradley, B. (1999). Some methodological issues in assessing attentional biases for threatening faces in anxiety: A replication study using a modified version of the probe detection task. *Behaviour Research & Therapy*, *37*(6), 595–604.

Niu, Y., Todd, R. M., & Anderson, A. K. (2012). Affective salience can reverse the effects of stimulus-driven salience on eye movements in complex scenes. *Frontiers in Psychology*, *3,* 336.

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, *171*, 308–311.

Nummenmaa, L., Hyönä, J., & Calvo, M. G. (2006). Eye movement assessment of selective attentional capture by emotional pictures. *Emotion*, *6*(2), 257–268.

Ohman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Emotion*, *130*(3), 466–478.

Olofsson, J. K., Nordin, S., Sequeira, H., & Polich, J. (2008). Affective picture processing: An integrative review of ERP findings. *Biological Psychology*, *77*, 247–265.

Onat, S., Açık, A., Schumann, F., & König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. *PloS One*, *9*(4), e93254.

Packard, M. G., & Cahill, L. (2001). Affective modulation of multiple memory systems. *Current Opinion in Neurobiology*, *11*(6), 752–756.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, *16*(2), 125–154.

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, *460*(7251), 94–97.

Pessoa, L. (2011). Reprint of: Emotion and cognition and the amygdala: From "what is it? to "what's to be done?" *Neuropsychologia*, *49*(4), 681–694.

Phelps, E. A. (2004). Human emotion and memory:

Interactions of the amygdala and hippocampal complex. *Current Opinion in Neurobiology*, *14*(2), 198–202.

Potter, M. (1975). Meaning in visual search. *Science*, *187*(4180), 965–966.

Pourtois, G., Grandjean, D., Sander, D., & Vuilleumier, P. (2004). Electrophysiological correlates of rapid spatial orienting towards fearful faces. *Cerebral Cortex*, *14*(6), 619–633.

Rayner, K., Smith, T. J., Malcolm, G. L., & Henderson, J. M. (2009). Eye movements and visual encoding during scene perception. *Psychological Science*, *20*(1), 6–10.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network*, *10*(4), 341–350.

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*(1A), 31–40.

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*(6), 852–877.

Sabatinelli, D., Lang, P. J., Bradley, M. M., Costa, V. D., & Keil, A. (2009). The timing of emotional discrimination in human amygdala and ventral visual cortex. *Journal of Neuroscience*, *29*(47), 14864–14868.

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences*, *14*(4), 303–316.

Schupp, H., Stockburger, J., Schmälzle, R., Bublatzky, F., Weike, A., & Hamm, A. O. (2008). Visual noise effects on emotion perception: Brain potentials and stimulus identification. *Neuroreport*, *19*(2), 167–171.

Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. *Journal of Vision*, *11*(5):9, 1–30, http://www.journalofvision.org/content/11/5/9, doi:10.1167/11.5.9. [PubMed] [Article]

Smith, N. K., Cacioppo, J. T., Larsen, J. T., & Chartrand, T. L. (2003). May I have your attention, please: Electrocortical responses to positive and negative stimuli. *Neuropsychologia*, *41*(2), 171–183.

Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4/, doi:10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659.

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*(12), 1857–1862.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5):5, 1–23, http://www.journalofvision.org/content/11/5/5, doi:10.1167/11.5.5. [PubMed] [Article]

Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica*, *135*(2), 77–99.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Tremblay, S., Saint-Aubin, J., & Jalbert, A. (2006). Rehearsal in serial memory for visual-spatial information: Evidence from eye movements. *Psychonomic Bulletin & Review*, *13*(3), 452–457.

Underwood, G., Foulsham, T., van Loon, E., & Underwood, J. (2005). Visual attention, visual saliency and eye movements during the inspection of natural scenes. In J. Mira & J. R. Alvarez (Eds.), *Artificial intelligence and knowledge engineering applications: a bioinspired approach* (pp. 459–468). Berlin: Springer-Verlag.

Võ, M. L., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, *9*(3):24, 1–15, http://www.journalofvision.org/content/9/3/24, doi:10.1167/9.3.24. [PubMed] [Article]

Võ, M. L., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, *10*(3):14, 1–13, http://www.journalofvision.org/content/10/3/14, doi:10.1167/10.3.14. [PubMed] [Article]

Võ, M. L., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Attention, Perception, & Psychophysics*, *73*(6), 1742–1753.

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain: An event-related fMRI study. *Neuron*, *30*(3), 829–841.

Whalen, P. J. (1998). Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science*, *7*, 177–188.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience*, *18*(1), 411–418.

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research*, *46*(8–9), 1520–1529.

Wright, R. D., & Ward, L. M. (2008). *Orienting of attention*. New York: Oxford University Press.

Wu, C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*, 1–13.

Appendix A

Semantic load in fixations' locations

Semantic maps were obtained by asking participants to circle regions determining emotional category of an image. The agreement among participants in labeling areas crucial for emotional meaning of the image varied between valence categories, with the highest agreement in case of negative (74%), lower in case of positive (65%), and the lowest in case of neutral images (56%). Observed variability in marking indicates that in the negative condition, the key object was easy to identify, while in neutral images, semantic areas were less conspicuous.

To investigate how differences in markings affected our results, we conducted additional analysis that took into account agreement in semantic labeling at the point of fixation. For each picture, raw values of semantic map (i.e., selections of semantically relevant regions averaged for all participants) on the first two fixations' locations were analyzed. Positive sample was derived by calculating mean value on the semantic map in the fixations' locations executed while viewing analyzed image. Negative sample was derived by calculating mean value on the same semantic map in the fixations' locations executed while viewing other images. (In this case sample was simply mean value.) Then, positive mean value was divided by negative mean value, and was dubbed normalized semantic load (Figure A1C). This approach is to some extent comparable to the approach involving ROI segmentation presented in the Results (Figure 3), as they are alternative ways of assessing influence of semantics on fixations. However, taking into account all raw values of the semantic map prevents its segmentation into ROIs. Thus, semantic load cannot be calculated in a similar manner to proportion of fixations (*NFP*), i.e., separately for purely semantic and common regions (Figure A1A). The closest reasonable comparison is to all ROIs representing semantics, which is purely semantic and common ROIs merged (Figure A1B).

The chance of fixating in combined semantically rich and common regions calculated using *NFP* method depended on noise level, $F(10, 180) = 73.0$, $p < 0.001$;

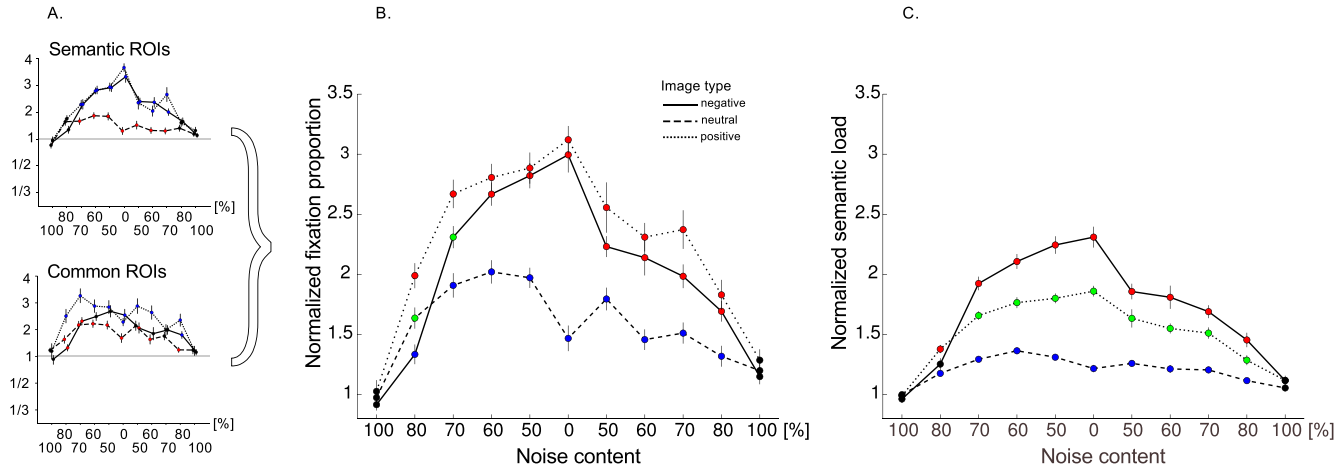

Figure A1. Comparison of methods based on calculating number of fixations in ROIs and semantic load in fixations' locations. (A) Normalized fixation proportions for semantic and common ROIs as presented in the Results section. (B) Normalized fixation proportion for combined semantic and common ROIs. (C) Normalized semantic load. The *x*-axis indicates percentage of noise content in an image arranged in the same order as presented in the experiment. Statistically significant differences between emotional conditions are encoded by color of dots (blue, red, and green) representing data points. If a pair differed significantly, corresponding dots have different colors. If a condition did not differ from any other, the dot is black. Error bars represent standard deviations.

valence, $F(2, 36) = 114.3$, $p < 0.001$; and interaction of those two factors, $F(20, 360) = 9.1$, $p < 0.001$. Trend analysis indicated that the relationship between noise level and fixation probability followed parabolic shape, $F(1, 18) = 341.0$, $p < 0.001$. Additionally, the interaction between noise level and valence was quadratic in nature, $F(1, 18) = 72.2$, $p < 0.001$, revealing that the curve describing the chance of fixating semantic region in consecutive noise conditions was more concave in case of negative and positive slides as compared to neutral slides.

Normalized semantic load on fixations' locations depended on noise level, $F(10, 180) = 111.2$, $p < 0.001$; valence, $F(2, 36) = 201.3$, $p < 0.001$; and their interaction, $F(20, 360) = 23.1$, $p < 0.001$. Trend analysis proved that the relationship between noise level and semantic load was parabolic, $F(1, 18) = 371.3$, $p < 0.001$. Interaction between noise level and valence was significant in both linear, $F(1, 18) = 57.9$, $p < 0.001$, and quadratic, $F(1, 18) = 153.9$, $p < 0.00$, trends, indicating that the curves describing the semantic load in consecutive noise levels differed between all valence conditions.

The curves obtained using normalized semantic load assume similar overall shapes, amid lower values, to those calculated using ROIs. Specifically, the difference between neutral and both emotional conditions, apparent while using ROIs, is maintained when agreement in semantic labeling is taken into account. However, a significant difference between negative and positive conditions becomes visible. This difference is reminiscent of the so called "negativity effect," which assumes that attention is engaged more strongly by negative than positive information. This is in contrast to the "emotionality effect," posing general attentional preference to emotional stimuli, regardless of whether negative or positive, as opposed to neutral ones (Humphrey et al., 2012). In eye-tracking studies, negativity effect was demonstrated either when more than few initial fixations were taken into account (Humphrey et al., 2012) or when an additional variable, like arousal of the emotional stimulus, was introduced to the fixation probability analysis (Niu et al., 2012). Conversely, once analysis was limited to the first fixation only, only a general emotionality effect was apparent. Therefore, it appears that the presence of either negativity or general emotionality effects is driven largely by particular calculation method, a phenomenon present also in our data. Specifically, using the ROI method we are detecting only general emotionality effect, while addition of agreement in labeling leads to emergence of the negativity effect. The influence of noise level on fixation probability and semantic load seems to be similar, as confirmed by the same effects in trend analysis. In conclusion, including labeling agreement in calculations changes relationship between valences but has no impact on noise effect.

Although the presented manner of calculating the results takes into account differences in labeling agreement, it does not allow for differentiation into separate purely semantic, purely visually salient, and common ROIs. Therefore, it is not optimal method for studying interaction between semantics and visual salience.

Appendix B

Areas under ROC curves for semantic and saliency maps

The *NFP* measure presented in the Results section is based on selecting ROIs, which allows to distinguish purely salient (visually salient ROIs), purely semantic (semantic ROIs), and overlapping regions (common ROIs). The benefit of such procedure is the possibility of assessing the interaction between semantics and visual saliency.

However, selecting semantic ROIs based on one arbitrary threshold may cause a bias in results. Therefore, we calculated also ROC curve, a more commonly used measure, which takes into account the entire map without setting one permanent threshold. We calculated the ROC measure separately for each participant and each experimental condition (valence × noise level) for visual saliency map and semantic map. Then, we calculated area under curve (AUC) for each ROC curve and conducted repeated measures ANOVA analysis with factors of valence (negative, neutral, positive) and noise (11 levels).

Comparison between ROC curves and *NFP* measures for three types of ROIs separately cannot be easily done, since the overlap of the semantic and saliency map cannot be explicitly controlled in the classic ROC curve. Therefore, we merged the results for common ROIs with results for both semantic and salient ROIs to obtain *NFP* data that can be directly compared with areas under the ROC curve (Figure B1A). This merger had significant impact on results for visually salient ROIs (Figure B1D), while results for semantic ROIs remained essentially unchanged (Figure B1B). Please note that the *NFP* measure is interpreted as odds, while ROC curve is probability, which makes them easily convertible. Both units are presented in the graphs.

The chance of fixating in combined purely semantic and common regions calculated using *NFP* method depended on noise level, $F(10, 180) = 73.0$, $p < 0.001$; valence, $F(2, 36) = 114.3$, $p < 0.001$; and their

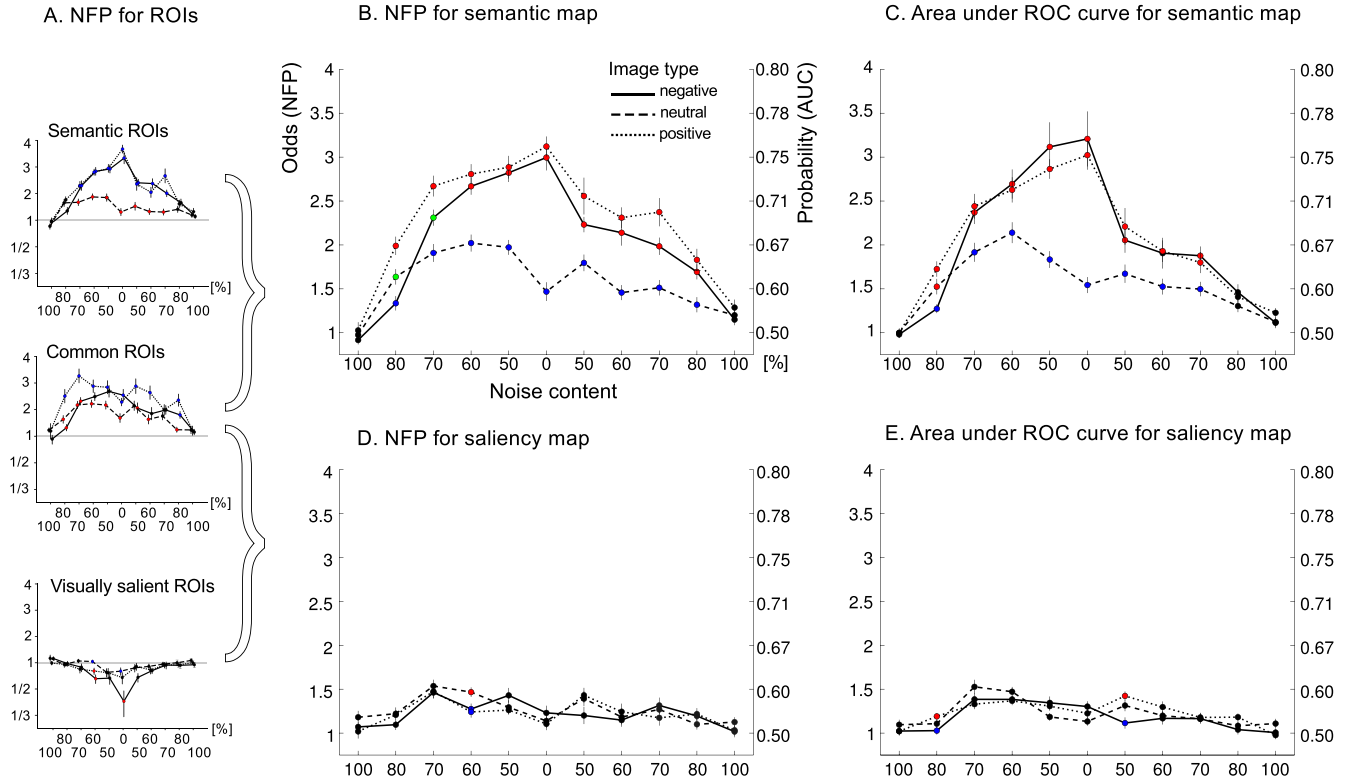

Figure B1. Comparison of results calculated using *NFP* and area under ROC curve. (A) Original results for semantic, common, and visually salient ROIs, presented in the Results section. (B) *NFP* measure for combined semantic and common ROIs. (C) Areas under ROC curves for semantic map. (D) *NFP* measure for combined visually salient and common ROIs. (E) Areas under ROC curves for visual saliency map. Probability—i.e., ROC original units, and odds, i.e., *NFP* original unit—are indicated on *y*-axis. On the *x*-axis, percentage of noise content in an image is indicated, arranged in the same order as presented in the experiment. Statistically significant differences between emotional conditions are encoded by color of dots (blue, red, and green) representing data points. If a pair differed significantly, corresponding dots have different colors. If a condition did not differ from any other, the dot is black. Error bars represent standard deviations.

interaction, $F(20, 360) = 9.1$, $p < 0.001$. Trend analysis proved that the relationship between noise level and fixation probability was parabolic, $F(1, 18) = 341.0$, $p < 0.001$. Additionally, the interaction between noise level and valence was quadratic in nature, $F(1, 18) = 72.2$, $p < 0.001$), revealing that the curve describing the chance of fixating semantic region in consecutive noise conditions was more concave in case of negative and positive slides as compared to neutral slides. The results calculated using the ROC method were virtually the same. The chance of fixating in semantic regions calculated using the ROC curve depended on noise level, $F(10, 180) = 102.8$, $p < 0.001$; valence, $F(2, 36) = 33.2$, $p < 0.001$; and interaction of those two factors, $F(20, 360) = 6.9$; $p < 0.001$. Relationship between noise level and fixation probability was parabolic, $F(1, 18) = 373.8$, $p < 0.001$, and interaction between noise level and valence was quadratic, $F(1, 18) = 38.4$, $p < 0.001$, indicating deepening of the curve in negative and positive conditions as compared to neutral.

The chance of fixating in combined visually salient and common regions calculated using the *NFP* method depended only on noise level, $F(10, 180) = 11.0$, $p < 0.001$. The relationship between the noise level and fixation probability was explained best by parabolic curve, $F(1, 18) = 28.2$, $p < 0.001$. In comparison, chance of fixating in the visually salient regions calculated using the ROC method depended on noise level, $F(10, 180) = 26.6$, $p < 0.001$, which was most prominent in the quadratic trend, $F(1, 18) = 118.1$, $p < 0.001$, but also on interaction of noise level with valence, $F(20, 360) = 2.2$, $p = 0.02$. This interaction, however, showed no regular tendency as both linear, quadratic, and cubic trends did not yield significant effects.

In summary, ROC and *NFP* measures produced very similar results both for the semantic map (Figure B1B, B1C) and for the visual saliency map (Figure B1D, B1E). Therefore, it seems that narrowing our analysis to ROIs rather than analyzing entire maps did not affect our results. The ROC method, although more widespread, does not allow for segmentation into ROIs and as such is less suitable for the purposes of our study.